METHOD FOR SEGMENTING NON-SEGMENTED TEXT USING SYNTACTIC PARSE

ABSTRACT OF THE DISCLOSURE

5   Embodiments of the present invention provide a method and apparatus for segmenting text by providing orthographic and inflectional variations to a syntactic parser. Under the present invention, possible segments are first identified in the sequence of characters. At least two of the identified segments overlap each other. For at least one of the segments, an alternative sequence of characters is identified. In some cases, this alternative sequence is formed through inflectional morphology, which identifies a different lexical form for a word identified by the segment. In some cases, the alternative sequence represents an orthographic variant of a word identified by the segment. The identified segments and the alternative segments are then passed to a syntactic analyzer, which produces one or more syntactic parses. The segments found in the resulting parses represent the segmentation of the input sequence of characters.